UC3
UC Curation Center

"Neighborhood Watch" for Repository Quality Assurance

National Neighborhood Watch Institute

Rev. 0.2 – 2011-09-08

Stephen Abrams
Patricia Cruse
John Kunze

University of California Curation Center
California Digital Library

Introduction

As information technology and resources become ever more central to all aspects of science, culture, education, and entertainment, the need for large-scale repositories and preservation programs for digital assets also increases. Before (and after) entrusting valuable digital assets to a repository for ongoing stewardship, it is important to have tools available to evaluate repository quality. A number of initiatives provide criteria and auditing procedures for repository trustworthiness, including [ISO 1636] and [TRAC]. These tools have been usefully applied to repositories in the context of self- and third-party-audits of trustworthiness; that is, used by repository managers to gauge the quality of their own systems and operational processes, or by external examiners to assert the level of conformance to established trustworthiness criteria. To date, however, there are few effective tools in the hands of second-parties, that is, repository customers, useful for making determinations of repository quality assurance.

Any quality assurance metrics should meet three important criteria:

- Objectivity.
- Repeatability.
- Independent verifiability.

Taken together, these ensure that a potential (or actual) repository customer can determine for herself/himself/itself an unambiguous measure of repository quality. (These criteria are familiar in the scholarly realm as they have historically formed the foundation for reputable academic discourse, peer review, and publication.) A fourth criterion is also useful in facilitating the widest possible adoption of a potential tool:

- Simplicity.

The technical sophistication and competency of repository customers varies widely, so in many cases a simple and minimally functional solution is often preferable to a more fully functioned, but complex tool.

An important initial metric is the determination of a repository's ability to maintain the bit-level integrity of managed digital assets. While not fully sufficient in and of itself to ensure the long-term usability of a digital asset, bit-level preservation is nevertheless a necessary foundation: if the bits are not available in authentic form then no further added-value preservation activities can be performed.

A useful analogy to deploy in this respect is the "neighborhood watch" [Watch]. In the same way that neighbors watch over neighbors for early detection of crime and other potentially harmful situations, repositories should support mechanisms to permit their customers to verify for themselves that their content is safe and accessible, and that the repository is meeting its baseline stewardship obligations.

Neighborhood watch

The essence of a repository neighborhood watch is simple: a repository provides access to managed assets via stable URLs; the repository's "neighbors" – that is, its customers – maintain lists of those URLs, each associated with a message digest known to be correct for the underlying asset. The customer can then determine on an objective, repeatable, and independently verifiable basis whether or not the repository is meeting its primary stewardship obligations. In essence, this is an augmented form of link checking. Decisions regarding periodicity and sampling coverage are a matter of local policy and mutual agreement between a repository and its customers.

(Another useful analogy for this type of activity is a file system "scrub", the process of verifying the bit-level integrity of all disk blocks, supported by many modern file systems such as [ZFS]. Similarly, a repository "scrub" verifies the bit-level integrity of all assets managed in that repository.)

Merritt Fixity Service

One specific tool that could be used as the basis for a repository neighborhood watch is the Merritt Fixity service [Fixity]. The Merritt repository operated by the University of California Curation Center (UC3) is based on the micro-services architectural paradigm in which the full range of repository is decomposed into a granular set of independent, but interoperable micro-services [Merritt, Micro]. Thus, the Fixity service can be deployed easily as a stand-alone RESTful web application.

The main data structure captures the following elements for each item registered with the service:

Name	***Type***	***Value***
URL	String	URL of the item in a repository.
Size	Integer	Veridical size, in octets.
Algorithm	Enum	Message digest algorithm: • Alder-32, CRC-32 • MD2, MD5 • SHA-1, SHA-256, SHA-384, SHA-512

Digest	String	Veridical message digest.
Verified	Date	Timestamp of last verification.
Status	Enum	Verification status: • Unverified • Verified • Size-mismatch • Digest-mismatch • Unavailable
LastSize	Integer	Last size, in octets.
LastDigest	String	Last message digest.

This structure provides all of the information necessary for the independent verification of a digital asset managed within a repository. New Item records should be established in the service at the time of asset submission to the repository, either for all such assets or a statistically meaningful subset. The Fixity service does not currently measure repository response time latency or retrieval throughput, but it could be instrumented to do so easily if performance-related metrics are also deemed significant.

Other tools similar in function to the Merritt Fixity service also could be deployed as the basis for a neighborhood watch, such as the UMIACS ACE (Audit Control Environment) [ACE] incorporated into the Chronopolis distributed repository federation [Chronopolis].

Conclusion

Repository users need effective tools for evaluating the reliability and trustworthiness of those repositories. A repository neighborhood watch permits users to verify independently and at will the bit-level integrity of digital assets managed in the repository. The Merritt Fixity service supports the necessary functions for an effective neighborhood watch.

Trust in a digital repository by its users is a valuable commodity, which should be difficult to win and easy to lose. Just as neighborhood reputation is not so much a matter of self-assertion – what you say about yourself is less important than what others say about you – the reputation of a repository is best reflected by the opinions of its customers. A repository neighborhood watch provides a way for repository users to "trust, but verify" [Verify], and thus be in a position to make realistic and empirically-based decisions regarding repository use.

UC Curation Center

The UC Curation Center is a creative partnership bringing together the resources and expertise of the California Digital Library (CDL), the 10 campuses of the University of California (UC), and external partners in the international digital curation community. UC3 provides innovative solutions for the long-term preservation and use of the University's valuable digital content.

References

[ACE] UMIACS, *ACE:Main*, 2011 <https://wiki.umiacs.umd.edu/adapt/index.php/Ace>.

[Chronopolis] SDSC, *Chronopolis:*

[Fixity] UC3, *Curation Fixity Service*, 2011 <http://www.cdlib.org/uc3/curation/fixity.htm>.

[ISO 16363] ISO/DIS 16363, *Space data and information transfer systems – Audit and certification of trustworthy digital repositories*, February 23, 2011. Available in draft form as CCSDS Red Book, October 2010 <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206521R1/Attachments/652x1r1.pdf >.

[Merritt] UC3, *Merritt*, 2011 <http://www.cdlib.org/uc3/merritt >.

[Micro] Stephen Abrams, Patricia Cruse, John Kunze, and David Minor, "Curation micro-services: A pipeline metaphor for repositories," *Journal of Digital Information* 12:2 (2011) <http://journals.tdl.org/jodi/article/view/1605>.

[TRAC] OCLC/CLR, *Trustworthy Repositories Audit & Certification: Criteria and Checklist*, Version 1.0, February 2007 <http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf>.

[UC3] UC3 , *University of California Curation Center*, 2011 <http://www.cdlib.org/uc3>.

[Verify] Wikipedia, *Trust, but verify*, 2011 <http://en.wikipedia.org/wiki/Trust,_but_verify>.

[Watch] National Neighborhood Watch Institute, *Home – National Neighborhood Watch Institute* <http://www.nnwi.org/>.

[ZFS] Wikipedia, *ZFS* <http://en.wikipedia.org/wiki/ZFS#Data_Integrity>.